RMD Open
Rheumatic & Musculoskeletal Diseases

ORIGINAL RESEARCH

Machine learning predicts response to TNF inhibitors in rheumatoid arthritis: results on the ESPOIR and ABIRISK cohorts

Vincent Bouget ,[1] Julien Duquesne,[1] Signe Hassler,[2,3] Paul-Henry Cournède,[4] Bruno Fautrel ,[5,6] Francis Guillemin ,[7] Marc Pallardy,[8,9] Philippe Broët,[2,3] Xavier Mariette ,[10] Samuel Bitoun [10]

To cite: Bouget V, Duquesne J, Hassler S, *et al*. Machine learning predicts response to TNF inhibitors in rheumatoid arthritis: results on the ESPOIR and ABIRISK cohorts. *RMD Open* 2022;**8**:e002442. doi:10.1136/rmdopen-2022-002442

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/rmdopen-2022-002442).

VB and JD are joint first authors.

Received 28 April 2022
Accepted 5 August 2022

Check for updates

© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Dr Samuel Bitoun;
samuel.bitoun@aphp.fr

ABSTRACT

Objectives Around 30% of patients with rheumatoid arthritis (RA) do not respond to tumour necrosis factor inhibitors (TNFi). We aimed to predict patient response to TNFi using machine learning on simple clinical and biological data.

Methods We used data from the RA ESPOIR cohort to train our models. The endpoints were the EULAR response and the change in Disease Activity Score (DAS28). We compared the performances of multiple models (linear regression, random forest, XGBoost and CatBoost) on the training set and cross-validated them using the area under the receiver operating characteristic curve (AUROC) or the mean squared error. The best model was then evaluated on a replication cohort (ABIRISK).

Results We included 161 patients from ESPOIR and 118 patients from ABIRISK. The key selected features were DAS28, lymphocytes, ALT (aspartate aminotransferase), neutrophils, age, weight, and smoking status. When predicting EULAR response, CatBoost achieved the best performances of the four tested models. It reached an AUROC of 0.72 (0.68–0.73) on the train set (ESPOIR). Better results were obtained on the train set when etanercept and monoclonal antibodies were analysed separately. On the test set (ABIRISK), these models respectively achieved on AUROC of 0.70 (0.57–0.82) and 0.71 (0.55–0.86). Two decision thresholds were tested. The first prioritised a high confidence in identifying responders and yielded a confidence up to 90% for predicting response. The second prioritised a high confidence in identifying inadequate responders and yielded a confidence up to 70% for predicting non-response. The change in DAS28 was predicted with an average error of 1.1 DAS28 points.

Conclusion The machine learning models developed allowed predicting patient response to TNFi exclusively using data available in clinical routine.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ There are no widely used clinical and/or biological factors that predict response to tumour necrosis factor inhibitors (TNFi) in patients with rheumatoid arthritis (RA).

WHAT THIS STUDY ADDS

⇒ We trained machine learning algorithms that use clinical and biological data to predict response to TNFi and validated them in a separate cohort.

⇒ We had a good overall prediction of TNFi response that was further improved when analysing etanercept and monoclonal TNFi separately.

⇒ We then set the decision thresholds of two different models to predict more accurately either response or non-response. We were thus able to accurately predict those two outcomes separately

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Accurately predicting non-response to TNFi could save time by directly using other targeted disease-modifying antirheumatic drugs.

⇒ Machine learning could improve management of patients with RA through predicting treatment response to TNFi.

INTRODUCTION

Rheumatoid arthritis (RA) is a complex disease, heterogeneous in its clinical presentation, severity and response to therapies. Targeted disease-modifying antirheumatic drugs (tDMARDs) are recommended for patients who do not respond to first-line methotrexate therapy. Thirteen drugs with different mechanisms of action may be considered for treatment, and tailoring treatment for an individual patient can be challenging.[1] Tumour necrosis factor inhibitors (TNFi) are frequently the first choice but unfortunately, around one-third of patients do not respond and would have benefited from another tDMARD.[2] There is a lack of simple markers to identify TNFi responders which

RMD Open: first published as 10.1136/rmdopen-2022-002442 on 23 August 2022. Downloaded from https://rmdopen.bmj.com on 7 March 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 1 Characteristics of the training set (ESPOIR) and validation set (ABIRISK) at last visit before treatment initiation

	ESPOIR	ABIRISK
Feature's name	**n=161**	**n=118**
Age, years	49 (13)	52 (13)
Female, n (%)	115 (71)	89 (76)
Weight, kg	69 (15)	75 (19)
Height, cm	166 (9)	167 (9)
Body mass index	24.9 (4.6)	25.6 (5.5)
Autoimmunity family history, n (%)	48 (30)	37 (32)
Ever smokers, n (%)	72 (45)	72 (61)
Current smokers, n (%)	28 (17)	33 (28)
Smoking cumulative dose, pack-year	8 (13)	15 (14)
Past pregnancy (among sex=female), n (%)	92 (74)	70 (79)
DAS28	4.6 (1.6)	4.4 (1.1)
CRP, mg/L	17 (27)	12 (16)
Erythrocyte sedimentation rate, mm	26 (23)	23 (19)
Creatinine, µmol/L	74 (17)	66 (14)
AST, UI/L	22 (8)	25 (13)
ALT, UI/L	22 (13)	27 (20)
White blood, cells/10^9	7.9 (2.6)	8.5 (3.2)
Neutrophils, cells/10^9	5.4 (2.4)	6.1 (3.2)
Lymphocytes, cells/10^9	1.7 (0.68)	2.1 (2.9)
Presence of anti-citrullinated protein antibody, n (%)	113 (70)	81 (70)
Presence of rheumatoid factor IgM	119 (74%)	78 (68%)
TNFi treatment sequences, N	N=208	N=118
Etanercept sequences, N (%)	100 (48)	68 (58)
Monoclonal anti-TNF antibodies sequences, N (%)	108 (52)	50 (42)
Adalimumab sequences, N (%)	80 (74)	39 (78)
Infliximab sequences, N (%)	17 (16)	11 (22)
Certolizumab sequences, N (%)	8 (7)	0 (0)
Golimumab sequences, N (%)	3 (3)	0 (0)
First TNFi line, N (%)	153 (74)	107 (91)
Non-responder imputation, N (%)	10 (4.8)	21 (18)
Responder to sequences, N (%)	122 (59)	72 (61)
Etanercept, N (%)	64 (64)	42 (62)
Monoclonal anti-TNF antibodies, N (%)	58 (54)	30 (60)
Co-treated with corticosteroids, N (%)	94 (45)	51 (43)
Co-treated with MTX, N (%)	152 (73)	64 (54)

Results are presented as follows: mean (SD) for continuous variables and amount (percentage) for binary variables.
ALT, aspartate aminotransferase; AST, alanine transaminase; CRP, C reactive protein; DAS28, disease activity score; MTX, methotrexate; TNFi, tumour necrosis factor inhibitors.

results in a trial–error process that is detrimental for the patient and costly for the healthcare system. Regarding RA management, predicting the therapeutic response or failure to a TNFi treatment prior to its initiation would be groundbreaking. Several studies have tried to identify biomarkers, but none of them have emerged as a reliable predictive factor of response.[2]

With the increasing amount of available data in medicine, new tools are required to extract information. Machine learning algorithms learn patterns from data and assume these will reproduce in the future. These algorithms identify patterns and rules without being explicitly programmed to do so, allowing unbiased discoveries. This is especially interesting in medicine to

RMD Open: first published as 10.1136/rmdopen-2022-002442 on 23 August 2022. Downloaded from https://rmdopen.bmj.com on 7 March 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 2 Result of the variable selection process for the prediction of the EULAR response

All TNFi	Etanercept	Monoclonal antibodies TNFi
DAS28	DAS28	DAS28
Age	Sex	Sex
Ever smoked	Ever smoked	Ever smoked
Weight	BMI	Weight
Lymphocytes		ESR
Neutrophils		
ALT		

The feature selection was run on the training set (ESPOIR).
ALT, alanine transaminase; BMI, body mass index; DAS28, disease activity score; ESR, erythrocyte sedimentation rate; TNFi, tumour necrosis factor inhibitors.

identify markers or combination of markers unknown so far by physicians. Machine learning is now widely used in healthcare, especially in radiology and oncology, whether it be for diagnosis, prognosis or treatment recommendation; and has proven itself very useful in assisting physicians for certain precise tasks.[3 4]

Recently, several studies have considered the increasing amount of available data and suggested that machine learning techniques could be clinically used to predict the TNFi therapeutic response.[5–8] These suggestions are backed by recent initiatives implementing machine learning models on a variety of RA datasets. Two of these initiatives use genetic data to predict patient response to TNFi.[9 10] Using this approach in clinical practice remains a challenge since genetic data are not widely available. Other initiatives[11 12] apply machine learning on clinical and biological data for similar objectives. However, these approaches often fail to describe how such models could be useful in clinical practice. Specifically, they do not study the impact of the decision threshold that transposes the probability yielded by the model into a binary therapeutic response. Defining this threshold requires a deep understanding of the clinical context to assess the clinical impact of such a tool. Unfortunately, this process is often poorly described. Another key element is the explanation of the model's predictions, as most studies do not show the impact of the variables in predicting the response to treatment. Underlining the key variables impacting the model predictions helps to build rheumatologists' confidence in the technique and promote further clinical use. Ultimately, most of these existing models might be of limited clinical use due to the absence of replication cohorts to validate the results.

This study builds and compares different machine learning models to predict the therapeutic TNFi response for patients with RA based on routinely available clinical and biological data. It is based on data from the ESPOIR cohort to train our models, and the results were validated on the ABIRISK cohort.

METHODS

Patients

We included patients from ESPOIR,[13] a French multicentric, longitudinal and prospective early arthritis cohort that constituted the training dataset. ESPOIR is an observational study that followed patients for 11 years; and where patients were treated according to each centre's clinical practice, without any modification for research. As validation cohort, we used ABIRISK,[14] a prospective study investigating the predictive factors of development of anti-drug antibodies in patients with RA treated with a first TNFi. Patients were followed until the end of the treatment and up to 18 months after the treatment initiation. Patients were included if they fulfilled the 2010 American College of Rheumatology/EULAR criteria and received at least one injection of TNFi. No restriction on disease activity, disease duration, failure of previous DMARDs nor co-medication was applied so that the selected population reflects the diversity observed in clinical practice. The decision to introduce a TNFi was left to the choice of the treating physician according to local practice. Patients stopping the TNFi treatment within 6 months after initiation due to pregnancy, surgery, poor compliance to the protocol or unknown reasons were excluded from the study population. Patients stopping the TNFi treatment within 6 months after initiation due to inefficacy or adverse events were considered as inadequate responders (non-responder imputation). Patients stopping the TNFi treatment within 6 months after initiation due to remission were considered as responders.

In the ESPOIR cohort, a patient could switch from one TNFi drug to another during the follow-up. Each treatment sequence was considered as independent from each other. To account for these multiple treatments, we included a binary variable equal to 0 for TNFi-naive patients at sequence initiation and 1 otherwise.

We performed the analysis in the population treated with any type of TNFi. We then divided the population into two sets: patients treated with etanercept and patients treated with monoclonal anti-TNF antibodies, anticipating a potential difference in the response to these two types of TNFi.

Variables

We have included all the daily practice clinical and biological variables that were available in both cohorts. They include demographics, smoking status, Disease Activity Score (DAS28), C reactive protein, disease characteristics, complete blood count parameters, and liver and kidney parameters as shown in table 1.

Endpoints

Two endpoints were considered, resulting in two distinct predictive tasks:

1. The primary endpoint was the prediction of the therapeutic response, defined as good or moderate EULAR response[15] assessed 12 months (±6 months due to yearly visits in the ESPOIR cohort) after treatment

RMD Open: first published as 10.1136/rmdopen-2022-002442 on 23 August 2022. Downloaded from https://rmdopen.bmj.com on 7 March 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

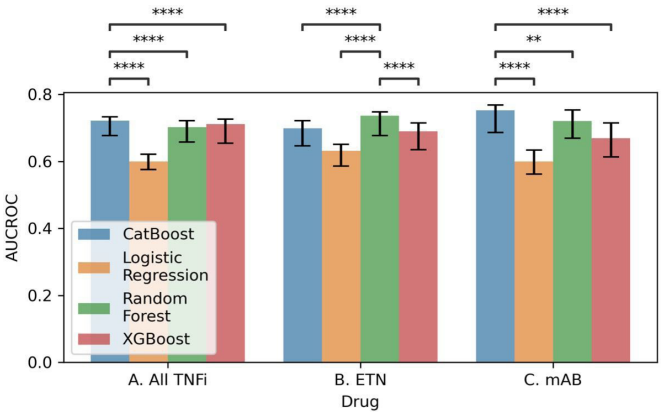


Figure 1 Performances of the models predicting the EULAR response calculated on the training set. Cross-validated AUROC of our models for each drug class on the training set with the 95% CI. The higher the AUROC, the better. Stars legend the p value ns: 5.00e-02<p≤1.00e+00 and ****p≤1.00e-04. AUROC, area under the receiver operating characteristic curve; ETN, etanercept; mAB, monoclonal antibodies; TNFi, tumour necrosis factor inhibitors.

initiation. This endpoint is binary (response vs inadequate response) and is analysed using the area under the receiver operating characteristic (ROC) curve (AUROC). A higher AUROC corresponds to a better model.

2. The secondary endpoint was the prediction of change in the erythrocyte sedimentation rate DAS28.[15] It is defined as the difference between DAS28 12 months after treatment initiation and DAS28 at treatment initiation. This difference is noted ΔDAS28 in the following. This endpoint is continuous and analysed using the mean squared error (MSE). A lower MSE corresponds to a better model.

Models

For both endpoints, the prediction pipeline is composed of consecutive blocks (online supplemental figure 1):

- A missing data imputation method.
- A variable selection process to select the predictive variables.
- A machine learning model predicting the outcome based on the predictive variables.

Four machine learning models were assessed: a linear regression model (logistic regression for the therapeutic response prediction and ridge regression for the ΔDAS28 prediction), a random forest model[16] and two gradient boosted trees models (XGBoost[17] and CatBoost[18]). Advantages and limits of these models are detailed in online supplemental methods.

Each model uses the variables available at the last visit before treatment initiation to predict the outcome. For the primary endpoint, the models output a score (between 0 and 100) which is interpreted as a probability for the patient to respond to TNFi. Ultimately, this score is compared with a decision threshold to obtain a binary prediction (response vs inadequate response). For the secondary endpoint, the models directly predict the DAS28 12 months after initiation.

The variables available in both cohorts are presented in table 1. Variable selection process and missing data imputation methods are presented in online supplemental methods.

Evaluation

The whole process (automated feature selection and model training) was evaluated using leave-one-out cross-validation on the training dataset. Only the best trained model was evaluated on the validation dataset to ensure the replication of the results. The metrics used to compare the models were the AUROC for the response classification (the higher, the better) and the MSE for the ΔDAS28 prediction (the lower, the better).

The prediction of the therapeutic response is a binary classification task and in such a case, classic epidemiological metrics exist. Specifically, for this task, we computed, to provide clinical perspective, the positive predictive value (PPV), the negative predictive value (NPV), the sensitivity and the specificity.

The models predict a probability for a patient to respond to TNFi. This probability is compared with a decision threshold to obtain the final binary outcome. The choice of this decision threshold requires to determine the best use case for the clinician. Two scenarios can be considered:

Table 3 Performances of the best models for the EULAR response prediction

Drug	Best model	Best missing value imputer	AUROC (train)	AUROC (validation)
Overall TNFi	CatBoost	MICE	0.72 (0.68 to 0.73)	Not evaluated since worse than drug-class-specific models on the training set
Etanercept	Random forest	Median	0.74 (0.68 to 0.75)	0.70 (0.57 to 0.82)
Monoclonal anti-TNF antibodies	CatBoost	MICE	0.74 (0.69 to 0.77)	0.71 (0.55 to 0.86)

The best model and best missing value imputer were selected on the training set (ESPOIR) using AUROC. The replication of the results was assessed on the validation set (ABIRISK). Numbers in brackets refer to 95% CIs.
AUROC, area under the receiver operating characteristic curve; MICE, Multiple Imputation by Chained Equations ; TNFi, tumour necrosis factor inhibitors.

RMD Open: first published as 10.1136/rmdopen-2022-002442 on 23 August 2022. Downloaded from https://rmdopen.bmj.com on 7 March 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

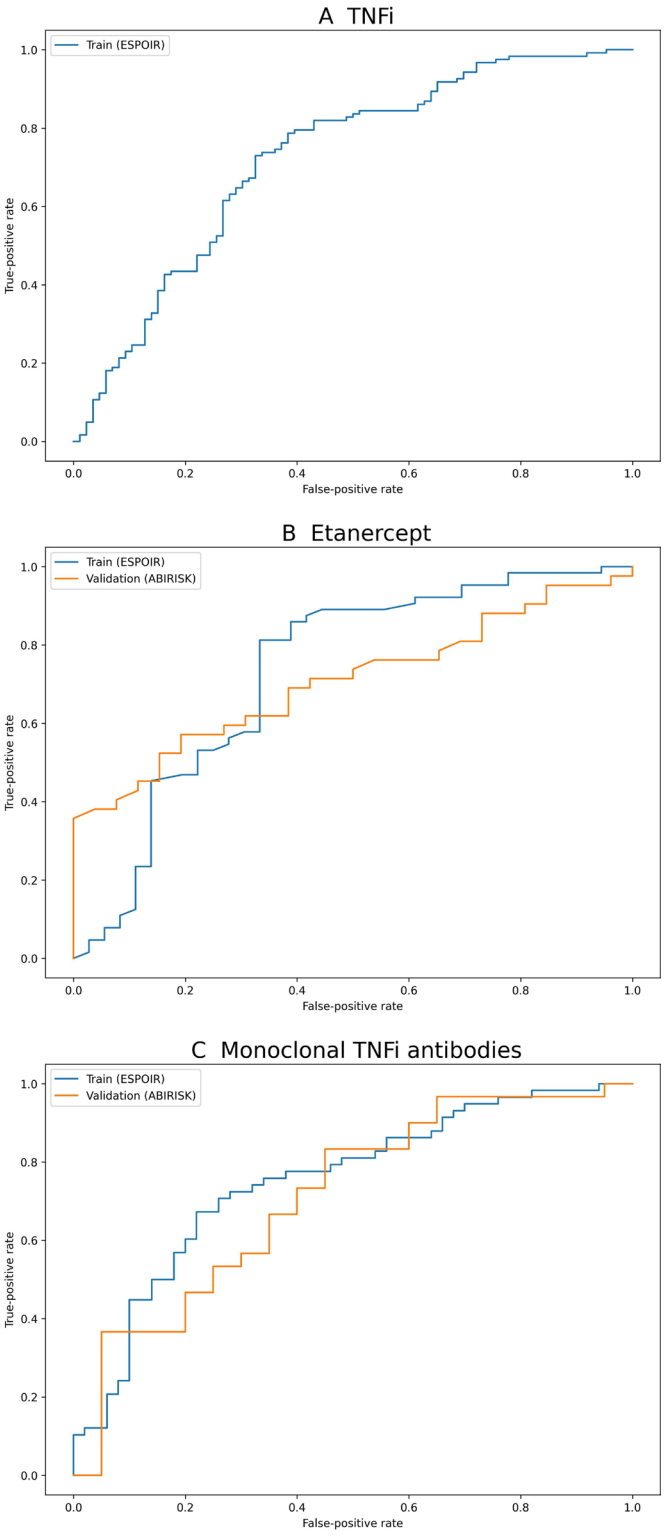


Figure 2 Performances of the best model predicting the EULAR response on the training and validation sets. ROC curves of the best models for the prediction of response to overall TNFi (A), etanercept (B) and monoclonal anti-TNF antibodies (C). ROC, receiver operating characteristic; TNFi, tumour necrosis factor inhibitors.

1. A high decision threshold guaranteeing the highest possible confidence when predicting a patient as responder to the treatment.
2. A low decision threshold guaranteeing the highest possible confidence when predicting a patient as inadequate responder to the treatment.

To choose the high threshold, we fixed a minimum PPV of 80% and selected the threshold that maximised the sensitivity on the cross-validated training set given this minimum PPV. To select the low threshold, we set a minimum value of 80% for the NPV. Given this minimum NPV, we selected the threshold that resulted in the best specificity on the cross-validated training set.

Details of the statistical methods for model comparison and CIs are detailed in online supplemental methods.

Explainability of the prediction of response to TNFi

Ultimately, we studied how the models yielded their decisions. One of the most popular packages to date to explain machine learning predictions is SHAP.[19] SHAP is based on the concept of Shapley value. A Shapley value is specific to a patient and to a characteristic. This value measures the weight of a patient characteristic on the patient outcome. A positive (resp. negative) Shapley value indicates a positive (resp. negative) influence on patient response to treatment. Higher Shapley values indicate stronger influences on patient response and vice versa.

We performed Shapley explanation on the ESPOIR and ABIRISK sets to obtain several visualisations for each model. We displayed explanation diagrams at the dataset level, plotting for each patient and each feature the contribution of this feature to the prediction.

RESULTS

Screening process

Among the patients from the ESPOIR cohort, 161 were included, of whom 95 initiated a treatment by etanercept and 96 initiated a treatment with an anti-TNF monoclonal antibody (in the ESPOIR cohort, 30 patients switched from one TNFi to another during the follow-up, resulting in more treatment sequences than the number of patients). In the ABIRISK cohort, 118 patients were included, of whom 68 initiated a treatment by etanercept and 50 initiated a treatment with adalimumab or infliximab.

The percentage of response to TNFi is 59% in the training set (ESPOIR) and 61% in the validation set (ABIRISK). The baseline characteristics of treatment sequences are displayed in table 1.

Results of the variable selection process

A backward feature selection process was applied to the variables presented in table 1 for each of the drug classes. The result of the feature selection process is presented in table 2 for the EULAR response and online supplemental table 2 for the ΔDAS28 prediction. Only these variables were used to train the models for the prediction of the therapeutic response and the prediction of the ΔDAS28.

RMD Open: first published as 10.1136/rmdopen-2022-002442 on 23 August 2022. Downloaded from https://rmdopen.bmj.com on 7 March 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 4 Metrics of interest regarding the prediction of the EULAR response

	Strategy 1 (high confidence in response)				Strategy 2 (high confidence in non-response)			
Drug	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV
Etanercept	60% (44% to 74%)	73% (55% to 89%)	78% (63% to 92%)	53% (36% to 69%)	95% (88% to 100%)	15% (4% to 30%)	64% (52% to 76%)	67% (20% to 100%)
Monoclonal anti-TNF antibodies	37% (20% to 55%)	95% (83% to 100%)	92% (73% to 100%)	50% (35% to 66%)	90% (78% to 100%)	40% (19% to 62%)	69% (54% to 84%)	73% (44% to 100%)

For the two strategies presented in methods, we display the metrics on the validation set (ABIRISK). Numbers in brackets refer to 95% CIs.
NPV, negative predictive value; PPV, positive predictive value; TNF, tumour necrosis factor.

Evaluation of the models

The four machine learning models were assessed for the prediction of the EULAR therapeutic response. First, all TNFi were considered. Then, the models were assessed on the etanercept and on the anti-TNF monoclonal antibodies groups. Results and comparison of the models are displayed in figure 1. Overall, CatBoost and random forest had the best performances, and the limited performances of the logistic regression compared with tree-based models suggest non-linear interaction effects between the variables.

Results were better on the training set when splitting all TNFi in etanercept and monoclonal TNFi antibodies, we then only evaluated the drug-class-specific models on the validation set. For each drug class, the model that performed the best on the training dataset was then assessed on the validation dataset (table 3).

ROC curves for the best models for each drug on both ESPOIR and ABIRISK are presented in figure 2. They showed that models had a good generalisability with overlapping curves between the training and validation datasets. Calibration curves are presented in online supplemental figures 5–7.

For the two decision strategies detailed in the evaluation section, we computed for each group of drugs the specificity, the sensitivity, the PPV and the NPV of the best model on the validation dataset(table 4). When the strategy maximising the confidence in the detection of the responders was used, we looked in particular at the specificity and the PPV of the model. We obtained a higher PPV in the monoclonal antibody group (92%) than in the etanercept group (78%). On the other hand, we focused on the sensitivity and the NPV for the strategy identifying inadequate responders confidently. Sensitivity replicated very well on the validation set and reached 90% whatever be the drug class.

Evaluation of the ΔDAS28 prediction models is detailed in online supplemental figure 3. Overall, the ridge regression model had the best performances suggesting limited non-linear interaction effects between the variables.

For each drug, the model that performed the best on the training set was then evaluated on the validation database (ABIRISK); the results are presented in online supplemental table 3 with their 95% CIs. The mean average error was computed on the validation set as it is easier to interpret. Overall, our models predicted the ΔDAS28 after treatment initiation with an error around 1.1 points of DAS28. This error should be compared with the 0.6 threshold, which is the minimum clinically relevant variation of the DAS28.

Explanation of the models

For the prediction of the therapeutic response, the models with the best performances on the validation dataset were the CatBoost and random forest models. To interpret these models, we computed the SHAP values on the concatenation of the training and validation datasets for each class of drug (online supplemental figure 2). As an example, high DAS28 was obviously associated with response to treatment, whereas a low value contributed to inadequate response for all TNFi. When plotted individually (Figure 3), non-linear interactions arise, especially for values above 5 for which the effect of DAS28 on response was limited. Biological variables such as ALT (alanine transaminase) and lymphocytes had a U shape relationship with prediction of response. Thus, high and low values were associated with response and mid-range with inadequate response (figure 3).

Explanation of the ΔDAS28 prediction models is detailed in online supplemental results.

DISCUSSION

This study establishes that machine learning models are efficient to assess the therapeutic TNFi response using exclusive data available in clinical routine. We obtained a good AUROC on the replication cohort. By underlining how each variable impacted the model's predictions, this study provides insight on how these predictions are made.

The first objective of the study was to assess the performances of a machine learning model based exclusively on clinical and biological data. It is worth noting that the results replicate properly between the training set (ESPOIR) and the validation set (ABIRISK). For the EULAR response prediction, the AUROC dropped by less than 0.05 points (etanercept and monoclonal anti-TNF antibodies). Interestingly, even though dividing

RMD Open: first published as 10.1136/rmdopen-2022-002442 on 23 August 2022. Downloaded from https://rmdopen.bmj.com on 7 March 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

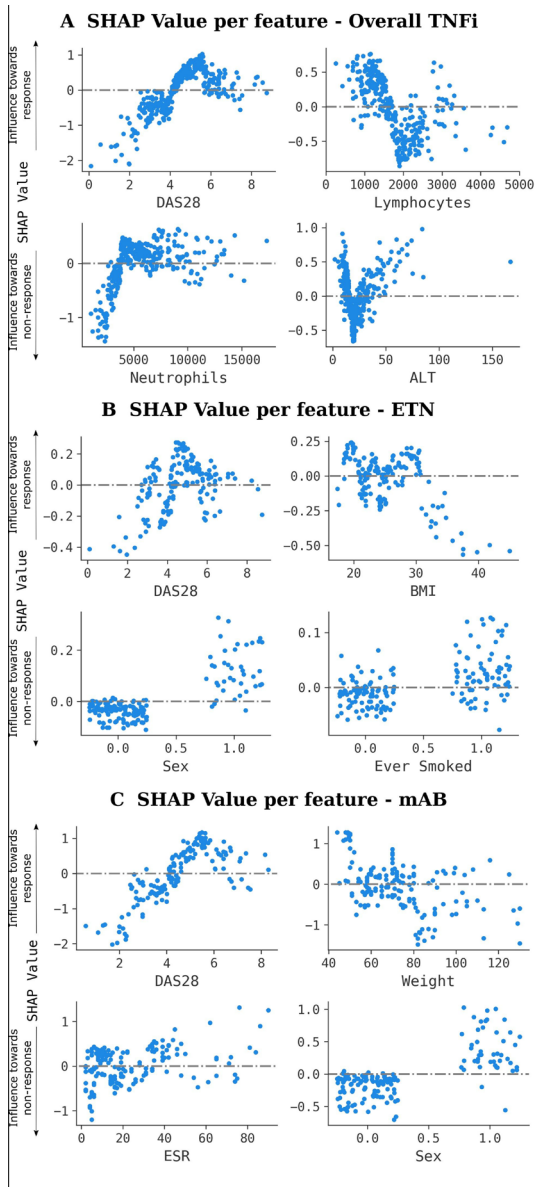


Figure 3 SHAP values of the best models for the prediction of the EULAR response to overall TNFi (A), etanercept (B) and monoclonal anti-TNF antibodies (C).The SHAP values are computed on a concatenation of the training and validation sets. Only the four most potent variables are displayed. Each dot represents a patient's data at treatment initiation and is placed on the y-axis according to its SHAP value and on the x-axis depending on the variable value. Positive (resp. negative) SHAP values influence the outcome towards a response (resp. inadequate response). This influence is proportional to the SHAP value. Female sex is encoded by 0. Jitter was added to binary variable to facilitate the reading. BMI, body mass index; DAS28, Disease Activity Score; ESR, erythrocyte sedimentation rate; ETN, etanercept; mAB, monoclonal antibodies; TNFi, tumour necrosis factor inhibitors.

by molecule groups leads to smaller study groups, which should decrease the prediction efficacy, the ΔDAS28 and the EULAR response prediction were more efficient when treatments were analysed separately (ie, etanercept and monoclonal anti-TNF antibodies) than together. This improvement in the results suggests differential mechanisms of action between the two classes of drug. This suggestion is strengthened by the observation that the variables used to predict response to etanercept or monoclonal anti-TNF antibodies are partially different or have different weight in the prediction (table 2 and online supplemental table 2).

The results of this study can be compared with similar initiatives using machine learning to predict treatment response to TNFi.[9–12] The prediction of the therapeutic response is the most frequent task; however, the only study providing a replication dataset and using the EULAR response as the primary endpoint is the publication of the winners of the DREAM RA Challenge.[9] Their model was developed using genomic data and reached an AUROC of 0.62. Numerically, our models seem to outperform theirs, especially when separating TNFi in groups of drugs. Nevertheless, large CIs do not allow us to conclude definitively. This comparison tends to confirm the potential of simple clinical and biological data for the identification of TNFi inadequate responders while extensive genetic data are not available in routine clinical practice. This remark is aligned with the conclusions of the DREAM RA challenge,[20] which noted that genetic data had limited effect on the final performance of the models.

The second objective of the study was to assess the potential of artificial intelligence-derived tools in clinical practice. Two different perspectives are offered to a clinician using these models: having the highest possible confidence of response when prescribing an anti-TNF or having the highest possible confidence when deciding to skip the anti-TNF drugs and directly prefer another tDMARD. The first scenario prioritises the PPV and the specificity of the algorithm, whereas the second prioritises the NPV and the sensitivity of the algorithm. We thus defined two decision thresholds to adapt the model to those two perspectives.

The first model brings the highest possible confidence in the detection of responders. It enables clinicians to prescribe TNFi with a 80% (etanercept) to 90% (monoclonal anti-TNF antibodies) confidence of response (PPV), which is much higher than the 60%–65% response rate observed in our cohorts and in the literature.[2] Using this algorithm and reaching this confidence would result in 50% (etanercept) to 75% (monoclonal anti-TNF antibodies) fewer patients treated with TNFi. Among these untreated patients, we would have missed 50% (etanercept and monoclonal anti-TNF antibodies) of responders (complementary of the NPV). These figures can be balanced given the large therapeutic arsenal of targeted treatments available in RA. Patients predicted as inadequate responders could easily be given another tDMARD.

The second model prioritises the confidence in the detection of inadequate responders. It enables clinicians to skip the TNFi and switch to another tDMARD with around 70% confidence of inadequate response (NPV). Reaching this level of confidence using our algorithm

RMD Open: first published as 10.1136/rmdopen-2022-002442 on 23 August 2022. Downloaded from https://rmdopen.bmj.com on 7 March 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

would result in 30%–35% less patients treated with TNFi. Among these untreated patients, we would have missed 30%–35% of responders (complementary of the NPV). The NPV of this strategy is quite satisfactory compared with the current clinical practice in which no clear markers are available.

The third objective of the study was to emphasise the explainability of the machine learning models. To provide a medical perspective, this study thoroughly details the explainability of the algorithm using the SHAP values. Our method was inspired by the successful application of SHAP values in oncology.[21] Several clinical studies assessed the influence of classic clinical variables on the response to TNFi.[2] Even if the conclusions on the influence of several variables are contradictory, a lot of consensus stands out and can be compared with the SHAP values of the models. In general, male sex, younger age and low body mass index (BMI) were associated with a response to TNFi, and smoking was associated with an inadequate response to TNFi. These results are consistent with the coefficients of the linear regression model for the prediction of the ΔDAS28 presented in online supplemental figure 4. As for the three classes of drugs (overall TNFi, etanercept and monoclonal anti-TNF antibodies), male sex, younger age and low BMI were associated with response, whereas smoking was associated with inadequate response. They are also consistent with the results of the EULAR response prediction presented (online supplemental figure 2) as young age and low weight are associated with response (positive SHAP values). The impact of the smoking status is unclear however, and the sex variable was not selected.

The SHAP values demonstrate that the models capture complex non-linear interactions between the features and the output. These non-linearities are particularly clear when plotting the Shapley values against DAS28 (figure 3). They also highlight turning points of influence between EULAR response and inadequate response; DAS28=4 appears as a clear frontier, for instance.

This study faces several limitations. The main limitation is the sample size which is rather small compared with data amount usually used in machine learning. The number of patients limits the accuracy of the algorithms and prevents from drawing very strong conclusions, given the spread of the CIs. Then, the use of data from two very different prospective cohorts results in a varying time between follow-ups. However, this variance highlights our model strengths since their performances are similar on those two very different cohorts. Modelling limitations can also be pointed out. In particular, for the prediction of the ΔDAS28, we only selected patients for which the DAS28 6 months after treatment initiation was available, which resulted in immortal time bias. This bias is not present in the prediction of the EULAR therapeutic response however, for which we used an inadequate response imputation rule. To improve the modelling of longitudinal data, Recurrent Neural Networks have been explored.[22 23] Although this modelling would complicate the insertion of this algorithm into a clinical environment, it could also result in a significant improvement.

To envision the use of this algorithm in clinical practice, we aim to design treatment strategy clinical trials. They would compare usual care with a decision based on the prediction of response by the algorithm.

CONCLUSION

In this study, we have demonstrated the ability of machine learning algorithms to predict response to TNFi using simple clinical and biological variables. Focusing on clinical use, we developed a model and assessed its performances in two scenarios, having a high confidence in either identifying TNFi responders or identifying TNFi inadequate responders. Both demonstrated interesting results compared with the current clinical practice and these algorithms pave the way to a personalised treatment strategy in RA.

Author affiliations
[1]Scienta Lab, Paris, France
[2]Sorbonne Université, INSERM UMR 959, Immunology-Immunopathology-Immunotherapy (i3), Assistance Publique Hôpitaux de Paris, Hôpital Pitié Salpêtrière, Paris, France
[3]CESP, INSERM UMR 1018, Paris-Saclay University, France, Villejuif, France
[4]CentraleSupélec Laboratory of Mathematics and Informatics for Systems Complexity, Gif-sur-Yvette, France
[5]Rheumatology Departement, Assistance Publique Hôpitaux de Paris, Groupe Hospitalier Pitié Salpêtrière, Paris, France
[6]Institut Pierre Louis d'épidémiologie et santé publique, Inserm UMRS 1136, équipe PEPITES (Pharmaco-épidémiologie et Évaluation des Soins), Paris, France
[7]Université de Lorraine, APEMAC, Vandoeuvre les Nancy, France
[8]INSERM UMR 996, Faculty of Pharmacy, Paris-Saclay University, Châtenay-Malabry, France
[9]ABIRISK (Anti-Biopharmaceutical Immunization: prediction and analysis of clinical relevance to minimize the RISK consortium), Innovative Medicines Initiative, Brussels, Belgium
[10]Rheumatology departement, Université Paris Saclay, Assistance Publique-Hôpitaux de Paris, Hôpital Bicêtre, INSERM UMR 1184, FHU CARE, Le Kremlin Bicêtre, France

Twitter Vincent Bouget @VincentBouget

Acknowledgements For the first 5 years of the ESPOIR cohort, an unrestricted grant from Merck Sharp and Dohme (MSD) was allocated. Two additional grants from INSERM were obtained to support part of the biological database. The French Society of Rheumatology, Pfizer, AbbVie, Lilly, and more recently, Fresenius and Biogen also supported the ESPOIR cohort study. We also wish to thank Nathalie Rincheval (Montpellier) who did expert monitoring and data management and all the investigators who recruited and followed the patients: F Berenbaum, Paris-Saint Antoine; M C Boissier, Paris-Bobigny; A Cantagrel, Toulouse; B Combe, Montpellier; M Dougados, Paris-Cochin; P Fardellone and P Boumier, Amiens; B Fautrel, Paris-La Pitié; R M Flipo, Lille; Ph Goupille, Tours; F Liote, Paris-Lariboisière; O Vittecoq, Rouen; X Mariette, Paris Bicetre; P Dieude, Paris Bichat; A Saraux, Brest; T Schaeverbeke, Bordeaux; J Sibilia, Strasbourg; as well as S Martin, Paris Bichat who did all the central dosages of CRP, IgA and IgM rheumatoid.We also thank investigators from the ABIRISK study for collecting and providing the data for this work.

Contributors SH and PB collected and cleaned the ABIRISK database. VB, JD, SB and XM conceptualised the study and performed the data request. VB and JD built the prediction models, analysed the data and prepared all figures. SB and XM assisted with interpretation of the data. P-HC provided expert guidance for all aspects of the study. VB, JD, SB and XM wrote the manuscript with support from P-HC, BF and MP. All authors reviewed the final manuscript. XM and SB are joint last authors and guarantor of this study.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

RMD Open: first published as 10.1136/rmdopen-2022-002442 on 23 August 2022. Downloaded from https://rmdopen.bmj.com on 7 March 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The protocol of the ESPOIR cohort study was approved in July 2002 by the ethical committee of Montpellier. The data are registered as ClinicalTrials.gov NCT03666091. The ABIRISK Study was approved by the IRB 'Paris Ile de France VII' under number 13-048 and by the French agency for drugs (ANSM) under number 2013-A01268-37. The ABIRISK Study was registered as study NCT02116504 by ClinicalTrials.gov. The respective scientific committees of both cohorts approved the use of the data for this study.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. Data used in this study are de-identified patient data and are not publicly available. They can be accessed after approval by the scientific committees of ABIRISK and ESPOIR cohorts.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Vincent Bouget http://orcid.org/0000-0002-9846-1942
Bruno Fautrel http://orcid.org/0000-0001-8845-4274
Francis Guillemin http://orcid.org/0000-0002-9860-7024
Xavier Mariette http://orcid.org/0000-0002-4244-5417
Samuel Bitoun http://orcid.org/0000-0003-0891-2269

REFERENCES

1 Smolen JS, Landewé RBM, Bijlsma JWJ, *et al*. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2019 update. *Ann Rheum Dis* 2020;79:685–99.
2 Wijbrandts CA, Tak PP. Prediction of response to targeted treatment in rheumatoid arthritis. *Mayo Clin Proc* 2017;92:1129–43.
3 Kingsmore KM, Puglisi CE, Grammer AC, *et al*. An introduction to machine learning and analysis of its use in rheumatic diseases. *Nat Rev Rheumatol* 2021;17:710–30.
4 Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018;2:719–31.
5 Hügle M, Omoumi P, van Laar JM, *et al*. Applied machine learning and artificial intelligence in rheumatology. *Rheumatol Adv Pract* 2020;4:rkaa005.
6 Nair N, Wilson AG. Can machine learning predict responses to TNF inhibitors? *Nat Rev Rheumatol* 2019;15:702–4.
7 Pandit A, Radstake TRDJ. Machine learning in rheumatology approaches the clinic. *Nat Rev Rheumatol* 2020;16:69–70.
8 Sutcliffe M, Radley G, Barton A. Personalized medicine in rheumatic diseases: how close are we to being able to use genetic biomarkers to predict response to TNF inhibitors? *Expert Rev Clin Immunol* 2020;16:389–96.
9 Guan Y, Zhang H, Quang D, *et al*. Machine learning to predict anti-tumor necrosis factor drug responses of rheumatoid arthritis patients by integrating clinical and genetic markers. *Arthritis Rheumatol* 2019;71:1987–96.
10 Tao W, Concepcion AN, Vianen M, *et al*. Multiomics and machine learning accurately predict clinical response to adalimumab and etanercept therapy in patients with rheumatoid arthritis. *Arthritis Rheumatol* 2021;73:212–22.
11 Koo BS, Eun S, Shin K. Explainable artificial intelligence for predicting remission in patients with rheumatoid arthritis treated with biologics. *Research square* 2021.
12 Lee S, Kang S, Eun Y, *et al*. Machine learning-based prediction model for responses of bDMARDs in patients with rheumatoid arthritis and ankylosing spondylitis. *Arthritis Res Ther* 2021;23:254.
13 Combe B, Benessiano J, Berenbaum F, *et al*. The ESPOIR cohort: a ten-year follow-up of early arthritis in France. *Joint Bone Spine* 2007;74:440–5.
14 Anon. Anti-Biopharmaceutical Immunization: Prediction and Analysis of Clinical Relevance to Minimize the RISK ABIRISK; 2019. abirisk.eu
15 Fransen J, van Riel PLCM, van RP. The disease activity score and the EULAR response criteria. *Clin Exp Rheumatol* 2005;23:S93–9.
16 Breiman L. *Mach learn*. 32. Random forests, 2001: 45–5.
17 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining*. San Francisco California USA: ACM, 2016: 785–94. https://dl.acm.org/doi/
18 Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. *ArXiv181011363 Cs Stat*, 2018. Available: http://arxiv.org/abs/1810.11363 [Accessed 15 Oct 2021].
19 Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: *Advances in neural information processing Systems. Vol 30*. Curran Associates, Inc, 2017. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
20 Sieberts SK, Zhu F, García-García J, *et al*. Crowdsourced assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis. *Nat Commun* 2016;7:12460.
21 Moncada-Torres A, van Maaren MC, Hendriks MP, *et al*. Explainable machine learning can outperform COX regression predictions and provide insights in breast cancer survival. *Sci Rep* 2021;11:6968.
22 Hügle M, Kalweit G, Hügle T. A Dynamic Deep Neural Network for Multimodal Clinical Data Analysis. In: Shaban-Nejad A, Michalowski M, Buckeridge DL, eds. *Explainable AI in healthcare and medicine. Studies in computational intelligence*. 914. Cham: Springer International Publishing, 2021: 79–92. http://link.springer.com/10.1007/978-3-030-53352-6_8
23 Norgeot B, Glicksberg BS, Trupin L, *et al*. Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis. *JAMA Netw Open* 2019;2:e190606.

RMD Open: first published as 10.1136/rmdopen-2022-002442 on 23 August 2022. Downloaded from https://rmdopen.bmj.com on 7 March 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.